INTERNATIONAL JOURNAL OF
RADIATION ONCOLOGY • BIOLOGY • PHYSICS
www.redjournal.org

PHYSICS CONTRIBUTION

Deep Learning-Based Automatic Assessment of Radiation Dermatitis in Patients With Nasopharyngeal Carcinoma

Ruiyan Ni, BSc,* Ta Zhou, PhD,* Ge Ren, PhD,* Yuanpeng Zhang, PhD,* Dongrong Yang, BSc,* Victor C.W. Tam, PgDPH,* Wan Shun Leung, PhD,* Hong Ge, MD, PhD,† Shara W.Y. Lee, PhD,* and Jing Cai, PhD*,‡

*Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China; †Department of Radiation Oncology, The Affiliated Cancer Hospital of Zhengzhou University, Zhengzhou, Henan, China; and ‡Research Institute for Intelligent Wearable Systems (RI-IWEAR), The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China

Received Aug 7, 2021; Accepted for publication Mar 7, 2022

Purpose: Radiation dermatitis (RD) is a common, unpleasant side effect of patients receiving radiation therapy. In clinical practice, the severity of RD is graded manually through visual inspection, which is labor intensive and often leads to large inter-rater variations. To overcome these shortcomings, this study aimed to develop an automatic RD assessment based on deep learning (DL) techniques that could efficiently assist the RD severity classification in clinical application.

Methods and Materials: A total of 1205 photographs of the head and neck region were collected from patients with nasopharyngeal carcinoma (NPC) undergoing radiation therapy. The severity of RD in these photographs was graded by 5 qualified assessors based on the Radiation Therapy Oncology Group guidance. An end-to-end RD grading framework was developed by combining a DL-based segmentation network and a DL-based RD severity classifier, which are used for segmenting the neck region from the camera-captured photographs and grading, respectively. U-Net was used for segmentation and another convolutional neural network classifier (DenseNet-121) was applied to RD severity classification. Dice similarity coefficient was used to evaluate the performance of segmentation. Severity classification was evaluated by several metrics, including overall accuracy, precision, recall, and F1 score.

Results: Results of segmentation showed that the averaged dice similarity coefficients were 91.2% and 90.8% for front and side view, respectively. For RD severity classification, the overall accuracy of test photographs was 83.0%. Our method accurately classified 90.5% of grade 0, 67.2% of grade 1, 93.8% of grade 2, and 100% of above grade 2 cases. The overall prediction performance was comparable with human assessors. There was no significant difference in accuracy when using manually or automatically segmented regions ($P = .683$).

Conclusions: We have successfully demonstrated a DL-based method for automatic assessment of RD severity in patients with NPC. This method holds great potential for efficient and effective assessing and monitoring of RD in patients with NPC.

Corresponding author: Jing Cai, PhD; E-mail: jing.cai@polyu.edu.hk

This research was partly supported by research grants from the Project of Strategic Importance Fund (P0035421) and the Project of RI-IWEAR fund (P0038684) from The Hong Kong Polytechnic University, and the Shenzhen-Hong Kong-Macau Science and Technology Program (Category C) (SGDX20201103095002019) and the Shenzhen Basic Research Program (R2021A067) from the Shenzhen Science and Technology Innovation Committee (SZSTI).

Disclosures: none.

Data sharing statement: Research data are not available at this time.

https://doi.org/10.1016/j.ijrobp.2022.03.011

Introduction

Acute radiation dermatitis (RD) is a common, unpleasant side effect when undergoing radiation therapy (RT) for patients with nasopharyngeal carcinoma (NPC). It may cause various complications without appropriate intervention, including diminished aesthetic appeal, a significant reduction in the patient's quality of life, and even death.[1] Studies on RD have been mainly focused on its pathogenic mechanism,[2] clinical manifestations and diagnosis,[3,4] prevention and treatment,[5] and prediction of changes.[6] It has been widely acknowledged that the early diagnosis, intervention, and management of RD requires the continuous monitoring of a patient's skin condition and prediction of its prognosis.[7] To guide the evaluation of RD severity in clinical practice, there are 3 commonly used RD grading criteria, including the Radiation Therapy Oncology Group (RTOG),[8] the National Cancer Institute Common Terminology Criteria for Adverse Events,[9] and the World Health Organization criteria.[10] Zenda et al[4] developed a grading atlas of RD for patients with NPC based on the Common Terminology Criteria for Adverse Events for improving the RD assessment quality in clinical practice. Many researchers have studied the effects of different drugs for RD prevention. Trolamine[11] and aloe vera[12] are 2 supportive ingredients that are commonly evaluated. Kawamura et al[6] developed a scoring system for acute RD prediction based on dosimetric and clinical factors. However, the current method of RD assessment for patients with NPC is based on visual inspection, which has 2 main weaknesses. First, the grading guidance, for example the RTOG grading criteria for skin toxicity, only uses descriptive words of RD symptoms. The visual assessment highly depends on the rater's experience and therefore it often leads to high interrater variations. Second, frequent RD assessment could be resource-demanding and may significantly increase clinicians' workload. It is therefore highly desirable to develop an automatic method for RD assessment to eliminate these shortcomings of the current RD assessment method and subsequently to improve RD management and patient overall treatment outcome.

Deep learning (DL) has been used as an automatic assessment tool for many skin diseases.[13-16] For instance, DL is commonly used for melanoma segmentation and classification.[17,18] Al Nazi and Abir[18] adopted a transfer learning approach to segment and detect melanoma automatically. The dermoscopic images were segmented via U-Net, a typical segmentation convolutional neural network (CNN) for medical imaging,[19] and then they did feature extraction on the region of interest (ROI) followed by melanoma classification. The mean dice score for the segmentation was 0.87 and classification accuracy reached 92%, which indicated that DL methods allow researchers to obtain satisfying results. Other similar studies also showed that models developed via DL methods had improved performance compared with those developed via traditional methods.[16] As for the prediction of RT toxic effects, some machine learning (ML)–based methods have been applied to patients with breast cancer[20-22] for accurately predicting RD severity and pain. Therefore, an automatic and reliable computer-aided system implemented with DL approaches would be an important RD severity evaluation tool for providing support and assistance to doctors.

The objective of this study was to develop an end-to-end RD grading framework, in which we input an RD photo and an RD severity grade could be outputted automatically, using DL techniques to facilitate the monitoring and management of RD in patients with NPC. To the best of our knowledge, our study is the first to introduce the DL technique for automatic RD assessment. An integrated 2-stage diagnostic model for neck region segmentation and multiple severity grade classification was developed via state-of-the-art DL techniques. This study can contribute in the following ways: (1) this model could effectively reduce the occurrence of misdiagnosis and reduce work burden; (2) with a similar prediction ability to human graders, it may eliminate the interrater variation caused by the subjectivity of human assessors and highly reduce the evaluation time; and (3) it could mimic the human assessors to accurately evaluate the RD severity but with more consistent assessment results and less evaluation time.

Methods and Materials

Data set

A data set of 1205 photographs of the head and neck region was built for patients with NPC undergoing RT, including 2 private data sets (740 and 373 photos, respectively) collected under Initial Review Board approved protocols from Queen Mary Hospital, Hong Kong SAR, China (ref. no. UW16-2002; HK cohort), and The Affiliated Cancer Hospital of Zhengzhou University, Zhengzhou, China (ref. no. 2019268; HN cohort) and 1 online public database[4] (92 photos). The photos were taken using digital cameras with a resolution of 96 dpi for patients' head and neck regions from 3 angles (front, left, and right oblique angles), as shown in Figure 1, according to a previous study.[4] To protect patient identification, the photos were trimmed to remove the face portion or patients' eyes were masked before the study.

For the 2 private data sets, the demographic and RT statistics of patients are shown in Table 1. For the HK cohort, the prescription dose was 66 to 70 Gy. Photos were taken at 8 time points for each patient, including 1 taken before the treatment and 7 taken weekly during the RT treatment course. For the HN cohort, the prescription dose was 64 to 70.4 Gy. Photos were taken every 5 fractions during the RT treatment course.

All photos were graded for RD severity by 5 qualified assessors based on the RTOG grading criteria for skin toxicity.[8] The qualified assessors who participated in the manual assessment of these photos evaluated all photos

Fig. 1. Protocol for photography. Three directions of photos were taken: right, front, and left oblique angles. A monochromatic background was used.

individually. All photos were randomly shuffled before grading to avoid any grading bias. For each photo, the most-voted grade among the assessors was used as the ground truth grade for model training. The individual grades of the assessors were also collected and used in the study. In total, the data set contained 493 photos of grade 0, 445 photos of grade 1, 213 photos of grade 2, and 54 photos with severity above grade 2. The photographs with severity above grade 2 were grouped together as "above grade 2" owing to their limited number. The split ratio for training, validation, and test set were 70%, 15%, and 15%, respectively. The example photographs of each RD grade are shown in Figure 2.

Study design

The proposed DL framework for RD severity diagnosis integrates a DL-based segmentation network for the ROI of the neck region and a DL-based classifier for RD grade evaluation from RD photographs. In this integrated framework, the neck ROIs were first segmented via U-Net, and after image postprocessing, the segmented ROIs were then inputted into a DenseNet-121[23]—inspired DL classifier. The final output would be 1 of the 4 RD categories: grade 0, grade 1, grade 2, and above grade 2. For further analysis and discussion, grade 0 and grade 1 were grouped as "mild" cases, and grade 2 and above grade 2 were grouped as "severe" cases.[24] According to the photo orientation, all images were divided into 2 subgroups, front view and side view (left and right views). Photographs with front and side views were trained under the same framework, and 2 segmentation models were generated and evaluated separately. During classification, 1 classification model was trained to grade all photos regardless of the view. For this task, we gave RD grades based on each RD photograph instead of each patient. Each RD photograph was individually handled and graded. The schematic diagram of the overall study design is shown in Figure 3.

Table 1 Demographic and radiation therapy statistics of patients

	HK cohort	HN cohort
Number of patients	27	31
Age (y)*	54 (15)	58 (18.5)
Sex (male:female)	18:9	25:6
Prescribed dose range	66-70 Gy	64-70.4 Gy
Overall fractions	33 or 35	30-33
Collection frequency	Weekly	Weekly

* Age is presented as median (interquartile range).

Automatic segmentation of RD regions via U-Net

ROI of the neck region was first manually delineated for all photos and was used to create binary masks of ROI. The photos and corresponding binary masks of ROI were then cropped into 512 × 512 pixels and inputted into a DL segmentation network based on a U-Net structure. Used for medical image segmentation tasks, U-Net, a fully convolutional network, can yield precise segmentation performance with limited training images.[19] It contained contracting and expansive paths. Four skip attentions were connected between the 2 paths for original shape recovery. The contracting path consisted of the repeated application of two 3 × 3 convolutional layers, each followed by a rectified linear unit (ReLU) and a 2 × 2 max-pooling operation for doubling feature channels. Each step in the expansive had feature channels with two 3 × 3 convolutions and an ReLU followed by a 2 × 2 convolution. At the last layer, a 1 × 1 convolution is used to map each 64-component feature vector to 1 class.

The segmented ROI masks were then postprocessed to further improve segmentation performance, using a maximum contour extraction and hole filling algorithm. The maximum contour extraction was used for small noise removal and the missing part inside ROI was filled up by the hole-filling algorithm. The segmented mask was overlaid on the original photograph to extract ROI (neck area) for further classification.

DL network for RD severity classification

A densely connected convolutional network, DenseNet-121,[23] was used for the RD severity classification. It has 4 dense blocks and 3 transition layers. Dense blocks contain 6, 12, 24, and 16 densely connected layers, respectively. In each dense block, each layer is connected to every other layer to preserve the feed-forward nature. The first part of the network consists of a 7 × 7 convolutional layer followed by a 3 × 3 max-pooling layer. Each dense block is composed

Fig. 2. Example photographs for each radiation dermatitis grade. Left to right, three directions of patients: right, front, and left views. Top to bottom, four radiation dermatitis grades: grade 0, grade 1, grade 2, and above grade 2.

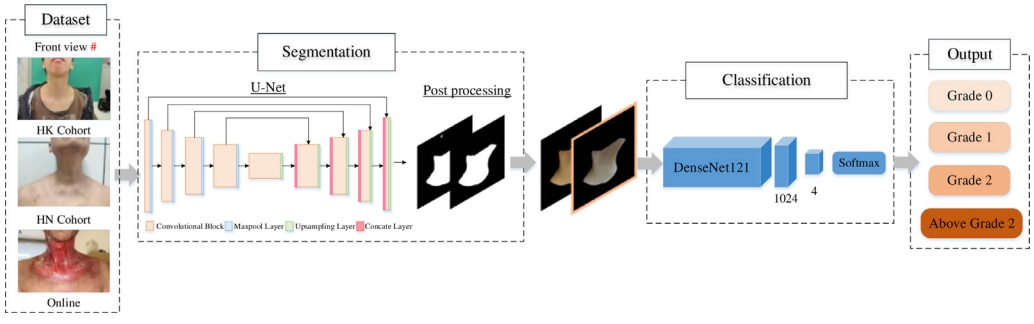

Fig. 3. Schematic diagram of the proposed integrated framework for region of interest segmentation and radiation dermatitis severity classification. Photo with the orange outline shows region of interest after color correction. #: During segmentation, photographs with front view and side view (left and right) were trained separately to generate 2 different segmentation models. Both views were trained together during classification. All photographs were trained under the same proposed framework.

of a batch normalization, a ReLu, a 3 × 3 convolutional layer, and a drop-out function. Between each dense block, the bottle neck layers are included, which contain 1 × 1 convolution to reduce the number of input feature maps and improve computational efficiency. A fully connected (FC) layer connected to the fourth dense block is used as a classification layer.

Data preparation

All the RD photographs were preprocessed before the model training. For the neck region segmentation, all RD photos and manually drawn binary masks were first center cropped into 512 × 512 pixels to remove background and focus on the ROI. For the RD severity classification part, the input data were resized into 256 × 256 pixels and then cropped into 224 × 224 pixels. The color variation of RD photos caused by different sources can reduce the classification accuracy. Therefore, color correction was conducted to ensure the color consistency among different RD photos.

Data balancing and augmentation

Our RD data set collected from participating hospitals and the online public database was imbalanced in terms of the number of photos in each RD severity grade. As the photos of different RD grades were highly imbalanced, the weighted class strategy was applied to the classification network. During classification training, a manual rescaling weight was assigned to each grade in the loss function, which was the cross entropy loss used in this study. The weight of grade a was calculated as the size of largest grade divided by the size of grade a. In this case, the input weight was a tensor with value of [1, 1.11, 2.31, 9.58]. Additionally, our current development of automatic diagnosis still suffered from the limited available RD photos. To deal with this problem, data augmentation was performed on the training set to manually increase the number of photographs and avoid the possible overfitting problem during the training process that may be caused by the small number of training data.[25] Augmentation techniques included the following methods: (1) random horizontal and vertical flipping (probability = 0.5); (2) random rotation within the range of −20° to 20°; and (3) random contrast adjustment (range was 0.5, 1.5). The data augmentation produced 3356 training photographs compared with the original unaugmented training data of 841 RD photos.

Color correction for RD photographs

To address the issue of photo color variations among different data sets, a color constancy algorithm, called the Grey-Edge algorithm, was used for color correction.[26] As an edge-based color constancy algorithm, Grey-Edge comprises zeroth-order methods, first-order, and higher-order methods. Based on the hypothesis that the average of the reflectance differences in a scene is achromatic,[26] the framework, by incorporating these color constancy methods, can be shown as:

$$\left(\int \left| \frac{\partial^n \mathbf{f}^\sigma(\mathrm{x})}{\partial \mathbf{x}^n} \right|^p d\mathrm{x} \right)^{\frac{1}{p}} = k\mathbf{e}^{n,p,\sigma} \quad (1)$$

where n indicates the order of method; $\mathbf{f}$ is the image value; $\mathbf{e}$ is the light source color; p is the Minkowski norm; and the scale of the local measurements is denoted by σ. Vectors are shown in bold. In this study, we set the Minkowski norm $p = 7$, $\sigma = 5$, and used the second-order method. As stated in the literature, best color constancy results may be obtained using this combination.[26]

Experimental setup

The whole data set was randomly split into a training data set with 841 photographs (345 of grade 0, 311 of grade 1, 149 of grade 2, and 36 of above grade 2), a validation data set with 182 photographs (74 of grade 0, 67 of grade 1, 32 of grade 2, and 9 of above grade 2), and a testing data set with 182 photographs (74 of grade 0, 67 of grade 1, 32 of grade 2, and 9 of above grade 2). To avoid the overfitting problem, data augmentation was conducted for the training set.

During the ROI segmentation, the models for front view and side view were trained individually for 90 epochs using U-Net. Binary cross entropy with logit loss was used as the loss function to minimize the difference between ground truth and network prediction. Root mean square propagation was used as the optimizer to update the model, with a learning rate of 1e-4, weight decay of 1e-8, and momentum of 0.9. The performance of the segmented ROI mask in comparison to manually delineated binary mask was evaluated by Dice similarity coefficient (DSC), which was computed using the following equation: DSC = (2 × the number of pixels in the overlap region of the ground truth and the predicted ROIs)/(the total number of pixels in the ground truth and the predicted ROIs).

For RD severity classification, the model was initialized with ImageNet pretrained weights and all layers were fine tuned. We used Adam optimizer, and cross entropy loss was chosen as the loss function. We set the hyper-parameters of the learning rate for feature extracting layers, the learning rate of classification layer, batch size, and decay to be 1e-5, 1e-4, 16, and 0.01, respectively. Fifty epochs were trained for the network to reach convergence. To quantitatively evaluate the predicting performance of RD severity classification, confusion matrix, and other evaluation metrics, including overall accuracy and precision, recall (also known as "sensitivity") and F1 score of each class were used. Precision quantified the number of positive predictions that belonged to the positive class. Recall was the true positive rate. F1 score was a single score that could balance both the concerns of precision and recall. Several experiments were conducted to evaluate the proposed method, including: (1) using manually segmented ROIs to test; (2) using photos without color

correction for testing; (3) only fine tuning the FC layer instead of all layers; and (4) using ResNet-50,[27] a commonly used deep CNN for image classification, with all layers fine-tuned. Wilcoxon signed rank test was performed to compare the proposed method with each experiment as the grades were not normally distributed. Additionally, the model performance was compared with the performance of each human assessor via the Wilcoxon signed rank test.

The implementation of the whole work was conducted on a computer with the following specifications: a CPU of Intel Core i9-10900KF @ 3.70GHz with 32 GB RAM and a Cuda-enabled GPU of NVIDIA GeForce GTX 1080 Ti. We implemented our work using Pytorch 1.7.0. All statistical analysis was performed with SPSS 25 (IBM, Chicago, IL). A significance threshold of P value <.05 was applied.

Results

Segmentation performance

The ROI (neck region) segmentation model achieved averaged DSCs of the testing data set as 88.32% and 88.30% for

Fig. 4. Sampled segmentation results for front view (first 2 rows) and side view (last 2 rows). For each subgroup, from left to right: original image, ground truth mask, automatically segmented result, overlapping between original image and segmentation result.

front and side view, respectively. After the image postprocessing, averaged DSC for the front view set was improved from 88.32% to 91.19% and the side view was improved from 88.30% to 90.84%. Figure 4 illustrates the segmentation results of some samples in both subgroups.

Multiclass classification performance

The 4-class classification performance of the proposed RD severity assessment framework was evaluated using 182 testing RD photographs randomly selected from all 3 cohorts. Overall training and testing accuracy were 97.4% and 83.0%, respectively. Their confusion matrices are presented in Figure 5a and 5b. More analysis of our model performance in the test data set is demonstrated in the following sections. First, more quantitative evaluations of the classification network testing performance are shown in Table 2 in terms of precision, recall, and F1 score. The results showed that all photos with above grade 2 were correctly predicted even with a small number of training samples, and there were no true negative samples for cases above grade 2. As for the classification performance between mild (<grade 2) and severe (≥grade 2) cases, 2 severe cases were wrongly predicted as mild cases and 4 mild cases were classified as severe. Furthermore, we trained a binary classifier of patients with mild and severe cases using the same data set. The overall testing accuracy of this binary classification model was 98.4% (179/182). The model successfully graded 97.9% (138/141) of mild cases and 100.0% (41/41) of severe cases.

Additionally, we compared our model performance with the individual evaluation results of human assessors. Figure 5c shows the grading accuracy of our automatic model and 5 qualified assessors. Our model had a higher accuracy than 2 out of 5 human assessors (assessor 2, assessor 3). Wilcoxon signed rank test was used to evaluate whether there was a significant difference between the model prediction results and each group of human labeling. The significance levels were 0.095, <0.001, 0.408, 0.039, and 0.655 for each human assessor, respectively. The model had a significantly better predicted performance compared with assessor 2, who was less experienced. For assessors 1, 3, and 5, there was no significant

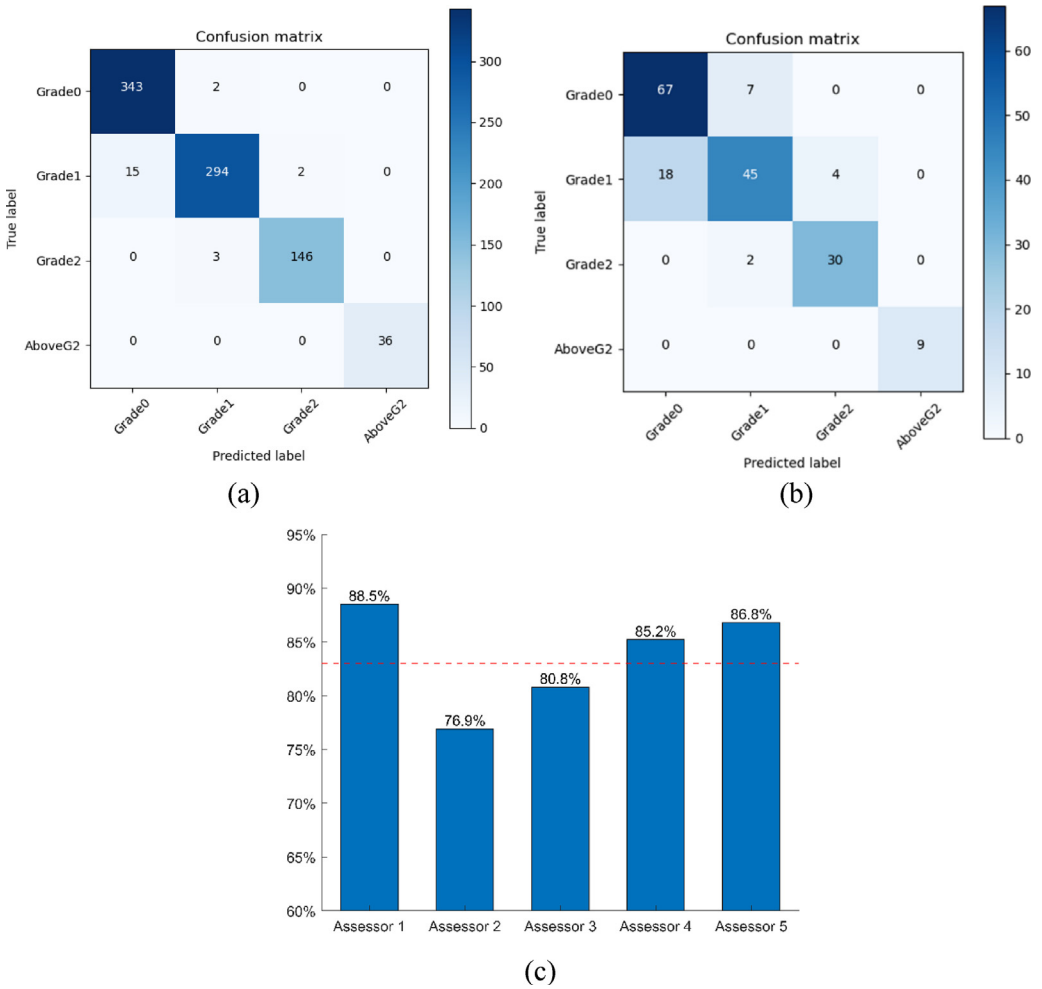

Fig. 5. (a) Confusion matrix of training data set. (b) Confusion matrix of testing data set. (c) Evaluation accuracy of human assessors and our proposed model. The red horizontal line indicates our model performance (83.0%).

Table 2 Evaluation metrics of radiation dermatitis severity classification

Class	Precision	Recall	F1 score
Grade 0	0.788	0.905	0.843
Grade 1	0.833	0.672	0.744
Grade 2	0.882	0.938	0.909
Above grade 2	1	1	1

difference between each of them and the DL model performance. Assessor 4 achieved a significantly better prediction than the model while the accuracy difference was only 2.2%. The reason may be that the assessor and our model wrongly labeled different cases, as they had only 5 false labels in common. Therefore, the overall results indicated that the proposed DL model could achieve comparable RD severity classification performance compared with human assessors and had significantly better prediction ability than the junior assessor.

Furthermore, some experiments were conducted to evaluate the effect of using different data processing and network fine-tuning methods, as shown in Table 3. First, the manually segmented ROIs were used as testing data instead of using automatically segmented ones. The accuracy difference was 1.1% and there was no significant difference between these 2 prediction results, which suggested that the auto-segmentation method was sufficient to delineate ROIs for severity classification. Second, compared with photos without color correction, using the color variation corrected photos in the proposed method could have a slightly better overall accuracy. Third, instead of fine tuning all layers, only FC was fine tuned in the third experiment, and a significant accuracy decrease (9.4%) could be observed. Last, ResNet-50 was used for grade classification with all layers fine-tuned during training. The overall accuracy was significantly lower (6.1%) than our proposed method.

Discussion

As a common and undesirable side effect of RT, radiation-induced dermatitis, which can significantly reduce a patient's quality of life, requires active monitoring and precise assessment of the skin condition to ensure appropriate and timely intervention. This study aimed to automatically assess RD severity via DL methods to overcome the limitations and difficulties of the current expert rating method, such as inconsistency and amount of time used. DL methods have been implemented in the medical field to deal with various clinical questions and they have achieved promising performance.[28,29] Automatic feature learning is the major difference between the DL approach and some traditional ML methods.[30] Instead of extracting handcrafted image features, CNNs as a powerful DL way can learn the useful representatives of inputs and generate predictions directly from the learned features. In this project, our end-to-end CNN framework contained 2 stages, the neck region segmentation and RD severity classification. The final outputs were 4 RD severity categories, grade 0, grade 1, grade 2, and above grade 2. Our model showed promising results that can be used in real clinical scenarios with further implementations.

Segmentation of ROIs (neck regions) was developed using U-Net. As a widely used and successful CNN for medical image segmentation, U-Net contains both down-sampling and up-sampling and makes input images reach their original sizes. Also, skip connections are implemented in U-Net to concatenate extracted features from down-sampling paths to up-sampling paths. The limited sample size is one of the major issues affecting the model performance. It is generally believed that data augmentation is a data-space solution to deal with the problem of limited data.[31] In the present study, we used geometric transformation and color space transformation to eliminate the positional and color biases present in the training data, and thus, more underground features could be learned by the DL model. Additionally, the current database is imbalanced in terms of the

Table 3 Evaluation for different testing experiments

		F1 score				
Experiment	Overall accuracy	Grade 0	Grade 1	Grade 2	Abovegrade 2	*P* value(compared with proposed method)
Manually segmented	84.1%	0.845	0.683	0.940	1	.683
Without color correction	79.7%	0.819	0.748	0.849	0.778	.123
Only fine-tuning FC layer	73.6%	0.795	0.682	0.687	0.727	.006*
ResNet-50	76.9%	0.806	0.723	0.789	0.737	.001*
Proposed method	**83.0%**	**0.843**	**0.744**	**0.909**	**1**	-

Abbreviation: FC = fully connected.
* $P < .05$.
The bolded row is our proposed method that used auto-segmented regions of interest with color correction, and all layers in the model were fine tuned. Other experiments are named after the differences between them and the proposed method.

number of RD photos in each RD severity class due to the real clinical situation. Although 80% to 90% of patients with head and neck cancer receiving RT develop RD, only approximately 20% of patients will experience severe RD.[32,33] Therefore, it was impossible to differentiate among the photos with above grade 2 based on the current data set and we applied the weighted class strategy to deal with the imbalanced classes.

Patients with RD with grade 0 to grade 1 (mild case) are the nonurgent cases who will receive the same treatment, whereas patients with grade 2 and above (severe case) are more urgent cases that require different and immediate medical treatments.[34,35] Therefore, in real clinical applications, evaluating whether the patients have developed severe RD symptoms is probably more important for proper and timely treatment. For this reason, the present study focused on achieving accurate RD severity assessment for differentiating mild and severe cases. Our results have demonstrated our current model was able to achieve this goal. In addition, patients with more severe RD conditions may require additional treatment. Therefore, classification within severe cases is another important diagnostic task. In the current study, we divided all the severe cases into grade 2 and above grade 2 as the patient number of these cases was limited. Despite the limited training photos, all above grade 2 photos were correctly predicted, and there was no confusion between these 2 classes.

The current model has limited performance in differentiating between grade 0 and grade 1, which is presumably due to the following reasons: (1) limited sample size; (2) the naturally subtle differences between these 2 classes, which may have led to noisy labels as the consensus grading of the assessors (used as ground truth) may not be the true grade; and (3) variations in photo taking despite closely following the protocol, such as lighting conditions and shooting angels, as they were mostly taken by radiotherapists who are not experienced photographers. This is a real-world challenging problem that needs further studies to address it. Therefore, in our future studies, we plan to: (1) collect more RD photos with more diverse RD grading levels; (2) develop the model to predict the probability of certain RD grade(s) instead of giving 1 particular grade, which mimics the actual clinical scenario of interrater variation; and (3) optimize photo- taking protocol, including but not limited to standardizing camera specifications, ambient light, and patient posture and position.

As the ultimate goal of our DL framework is to assist and partially replace human work in a more efficient way, our results demonstrated the efficiency and effectiveness of the proposed method. First, our model achieved a clinically acceptable performance. Over 80% of the DL model prediction agreed with the clinician and can be accepted by clinical standard. In particular, the model for classifying grade 2 and/or above RD performed exceptionally well, with 90% agreement with the clinician (ie, less than 10% need to be adjusted). For classifying grade 1 RD, which is challenging even for clinicians, the DL model achieved a stable and acceptable prediction success rate of around 70%, that is, around 30% needed to be adjusted. Compared with the 5 human assessors in this study, our model achieved higher accuracy than most assessors, indicating that our model can perform better than human assessors in most cases (~80%) and could be used as a good reference in real clinical practice. Nevertheless, we recognize that some senior assessors may be able to achieve higher accuracy than our model. Second, we evaluated the time efficiency of the DL model. It took only 0.02 seconds to grade 1 RD photo, compared with around 30 seconds for the human assessors. For photos with ambiguous RD presentations, human assessors may take up to 2 minutes to examine the photo thoroughly and decide on the grading. The DL method is much more efficient than manual assessment, with an increased evaluation efficiency of about 1500~6000 times.

Though DL approaches have been commonly used in many aspects of skin cancer, including skin lesion segmentation,[36,37] skin cancer detection,[38] and classification,[39] few studies have applied DL methods to facilitate RD diagnosis and treatment, especially for patients with NPC. Some studies have demonstrated the ability of DL- and ML-based methods for prediction of RT toxic effects in patients with breast cancer. Reddy et al[20] used an ML-based precision oncology approach to develop accurate prediction models for moist desquamation, grade $\geq$2 RD, and grade $\geq$2 breast pain in patients with breast cancer. Another breast cancer RD severity prediction model was developed by Lim and Joseph[21] based on 3-dimensional dose images and RT fractionation data. Their model achieved an area under the curve of 0.80 and 0.81 on the training and validation data set for discrimination between RD grade $\leq$1 and grade $\geq$2. Quantitative thermal imaging biomarkers were also used in the supervised ML method for RD early prediction, and the study found that the early thermal markers after 5 RT fractions could be predictive for RD severity in breast RT.[22] As visual inspection of skin conditions is the classical approach in clinical usage, our study exploited the DL method to directly diagnose RD severity via patients' skin photographs, which mimicked human visual assessment to provide the accurate RD classification.

As the first attempt to solve this real-world challenging problem via DL methods, the current study had some limitations. First, our sample size was limited, especially for cases above grade 2. Second, the classification results of grade 0 and grade 1 were relatively poor. Therefore, in following studies, we will further address these problems by: (1) enlarging our RD data set, including collecting more photos, especially photos of more severe RD, and diversifying patients' demographic and prescription information; (2) eliminating the noisy labels caused by human assessors, which may be done by using cross-validation of multiple assessors and multiple rounds to reduce the human discrimination; and (3) developing the model to predict the probability of RD grade instead of giving a particular grade, which may make the model become more practical.

Conclusions

In this study, a DL-based end-to-end framework including neck region auto-segmentation and RD grading classification was developed for automatic RD severity assessment in patients with NPC. Our model presents fast and accurate RD evaluation of patients' photographs with variable cohorts. Experimental results show that it has a comparable predicting accuracy to human assessors and high efficiency. Focusing on dealing with this real-world challenging problem, our work holds great potential for efficient and effective assessment and monitoring of RD for patients with NPC in clinical applications and thus improving patients' quality of life.

References

1. Singh M, Alavi A, Wong R, Akita S. Radiodermatitis: A review of our current understanding. *Am J Clin Dermatol* 2016;17:277–292.
2. Rosenthal A, Israilevich R, Moy R. Management of acute radiation dermatitis: A review of the literature and proposal for treatment algorithm. *J Am Acad Dermatol* 2019;81:558–567.
3. Huang CJ, Hou MF, Luo KH, et al. RTOG, CTCAE and WHO criteria for acute radiation dermatitis correlate with cutaneous blood flow measurements. *Breast* 2015;24:230–236.
4. Zenda S, Ota Y, Tachibana H, et al. A prospective picture collection study for a grading atlas of radiation dermatitis for clinical trials in head-and-neck cancer patients. *J Radiat Res* 2016;57:301–306.
5. Glover D, Harmer V. Radiotherapy-induced skin reactions: assessment and management. *Br J Nurs* 2014;23:S28. , S30-5.
6. Kawamura M, Yoshimura M, Asada H, Nakamura M, Matsuo Y, Mizowaki T. A scoring system predicting acute radiation dermatitis in patients with head and neck cancer treated with intensity-modulated radiotherapy. *Radiat Oncol* 2019;14.
7. Leventhal J, Young MR. Radiation dermatitis: Recognition, prevention, and management. *Oncology-Ny* 2017;31:885–899.
8. Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). *Int J Radiat Oncol* 1995;31:1341–1346.
9. Trotti A, Colevas AD, Setser A, et al. CTCAE v3.0: Development of a comprehensive grading system for the adverse effects of cancer treatment. *Semin Radiat Oncol* 2003;13:176–181.
10. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981;47:207–214.
11. Abbas H, Bensadoun RJ. Trolamine emulsion for the prevention of radiation dermatitis in patients with squamous cell carcinoma of the head and neck. *Support Care Cancer* 2012;20:185–190.
12. Haddad P, Amouzgar-Hashemi F, Samsami S, Chinichian S, Oghabian MA. Aloe vera for prevention of radiation-induced dermatitis: A self-controlled clinical trial. *Curr Oncol* 2013;20:E345–E348.
13. Sikkandar MY, Alrasheadi BA, Prakash NB, Hemalakshmi GR, Mohanarathinam A, Shankar K. Deep learning based an automated skin lesion segmentation and intelligent classification model. *J Amb Intel Hum Comp* 2021;12:3245–3255.
14. Li LF, Wang X, Hu WJ, Xiong NN, Du YX, Li BS. Deep learning in skin disease image recognition: A review. *IEEE Access* 2020;8:208264–208280.
15. Zhao XY, Wu X, Li FF, et al. The application of deep learning in the risk grading of skin tumors for patients using clinical images. *J Med Syst* 2019;43.
16. Baig R, Bibi M, Hamid A, Kausar S, Khalid S. Deep learning approaches toward skin lesion segmentation and classification from dermoscopic images - a review. *Curr Med Imaging* 2020;16:513–533.
17. Al-Masni MA, Kim DH, Kim TS. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Comput Meth Prog Bio* 2020;190.
18. Al Nazi Z, Abir TA. *Automatic Skin Lesion Segmentation and Melanoma Detection: Transfer Learning Approach With U-Net and DCNN-SVM*. Springer, Singapore; 2020:371–381.
19. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. *Lect Notes Comput Sc* 2015;9351:234–241.
20. Reddy J, Lindsay W, Berlind C, et al. Applying a machine learning approach to predict acute toxicities during radiation for breast cancer patients. *Int J Radiat Oncol Biol Phys* 2018;102:S59.
21. Lim A, Joseph KJ. Predicting radiation-adverse effects using three-dimensional dose and fractionation data: Radiation dermatitis. *Int J Radiat Oncol* 2019;105:E130.
22. Saednia K, Tabbarah S, Lagree A, et al. Quantitative thermal imaging biomarkers to detect acute skin toxicity from breast radiation therapy using supervised machine learning. *Int J Radiat Oncol* 2020;106:1071–1083.
23. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proc Cvpr Ieee* 2017;2261–2269.
24. Robijns J, Lodewijckx J, Claes S, et al. Photobiomodulation therapy for the prevention of acute radiation dermatitis in head and neck cancer patients (DERMISHEAD trial). *Radiother Oncol* 2021;158:268–275.
25. Jiao ZC, Gao XB, Wang Y, Li J. A deep feature based framework for breast masses classification. *Neurocomputing* 2016;197:221–231.
26. Van de Weijer J, Gevers T, Gijsenij A. Edge-based color constancy. *Ieee T Image Process* 2007;16:2207–2214.
27. He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. *Proc Cvpr Ieee* 2016;770–778.
28. Jia X, Ren L, Cai J. Clinical implementation of AI technologies will require interpretable AI models. *Med Phys* 2020;47:1–4.
29. Xing L, Krupinski EA, Cai J. Artificial intelligence will soon change the landscape of medical physics research and practice. *Med Phys* 2018;45:1791–1793.
30. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–444.
31. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data-Ger* 2019;6.
32. Chan RJ, Larsen E, Chan P. Re-examining the evidence in radiation dermatitis management literature: An overview and a critical appraisal of systematic reviews. *Int J Radiat Oncol Biol Phys* 2012;84:e357–e362.
33. Ferreira EB, Vasques CI, Gadia R, et al. Topical interventions to prevent acute radiation dermatitis in head and neck cancer patients: A systematic review. *Support Care Cancer* 2017;25:1001–1011.
34. Mendelsohn FA, Divino CM, Reis ED, Kerstein MD. Wound care after radiation therapy. *Adv Skin Wound Care* 2002;15:216–224.
35. Hymes SR, Strom EA, Fife C. Radiation dermatitis: Clinical presentation, pathophysiology, and treatment 2006. *J Am Acad Dermatol* 2006;54:28–46.
36. Jafari MH, Karimi N, Nasr-Esfahani E, et al. Skin lesion segmentation in clinical images using deep learning. *Int C Patt Recog* 2016;337–342.
37. Vesal S, Ravikumar N, Maier A. SkinNet: A deep learning framework for skin lesion segmentation. *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)* 2018, 1–3. https://doi.org/10.1109/NSSMIC.2018.8824732.
38. Dascalu A, David EO. Skin cancer detection by deep learning and sound analysis algorithms: A prospective clinical study of an elementary dermoscope. *Ebiomedicine* 2019;43:107–113.
39. Hosny KM, Kassem MA, Foaud MM. Skin cancer classification using deep learning and transfer learning. *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)* 2018, 90–93. https://doi.org/10.1109/CIBEC.2018.8641762.